Penalized ordinal logistic regression using cumulative logits

Clémence Karmann, Anne Gégout-Petit, Aurélie Muller-Gueudin

Penalized ordinal logistic regression using cumulative logits

Clémence Karmann, Anne Gégout, Aurélie Gueudin

Inria Nancy Grand Est - BIGS

clemence.karmann@inria.fr

Journée scientifique FCH :
"Méthodes et modèles pour comprendre les réseaux biologiques"
January 15th 2018

Introduction

- Analyze links between a variable (response) and covariates

Introduction

- Analyze links between a variable (response) and covariates
- Variable selection problem: identify relevant covariates

Introduction

- Analyze links between a variable (response) and covariates
- Variable selection problem: identify relevant covariates
- Ordinal response

Overview

1. Cumulative logit model
   - Generalities
   - Coefficients
2. Estimation, inference
   - Lasso estimation of the coefficients β
   - Penalty parameter and variable selection
3. Simulation studies
4. Application to network inference of zero-inflated data

1. Cumulative logit model
   - Generalities
   - Coefficients
2. Estimation, inference
   - Lasso estimation of the coefficients β
   - Penalty parameter and variable selection
3. Simulation studies
4. Application to network inference of zero-inflated data

Multi-class regression: cumulative logit model

Idea

Generalization of the logistic regression for a response Y with $K > 2$ ordered categories.

Multi-class regression: cumulative logit model

Idea

Generalization of the logistic regression for a response Y with $K > 2$ ordered categories.

If we have p covariates $X_1, \ldots, X_p$, we model $p_\beta^j(x) := \mathbb{P}_\beta(Y \leq j | X = x)$ for $j = 1, \ldots, K-1$, by:

$$\text{logit } p_\beta^j(x) = \alpha_j + \beta_1 x_1 + \ldots + \beta_p x_p,$$

i.e.:

$$p_\beta^j(x) = \frac{\exp(\alpha_j + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\alpha_j + \beta_1 x_1 + \ldots + \beta_p x_p)}$$

Coefficients

$$\boxed{\begin{array}{l}\text{Pour } j = 1, 2, \ldots, K-1 : \\ \text{logit } \mathbb{P}_\beta(Y \leq j | X = x) = \alpha_j + \beta_1 x_1 + \ldots + \beta_p x_p\end{array}}$$

- Coefficients $(\beta_i)_{i=1}^p$ do not depend on j.

Coefficients

$$\boxed{\begin{array}{l}\text{Pour } j = 1, 2, \ldots, K-1 : \\ \text{logit } \mathbb{P}_\beta(Y \leq j | X = x) = \alpha_j + \beta_1 x_1 + \ldots + \beta_p x_p\end{array}}$$

- Coefficients $(\beta_i)_{i=1}^{p}$ do not depend on j.
- $\alpha_1 < \alpha_2 < \ldots < \alpha_{K-1}$.

Coefficients

$$\boxed{\begin{array}{l}\text{Pour } j = 1, 2, \ldots, K-1 : \\ \text{logit } \mathbb{P}_\beta(Y \leq j | X = x) = \alpha_j + \beta_1 x_1 + \ldots + \beta_p x_p\end{array}}$$

- Coefficients $(\beta_i)_{i=1}^p$ do not depend on j.
- $\alpha_1 < \alpha_2 < \ldots < \alpha_{K-1}$.
- As $p_\beta^K(x) = 1$, there are $K - 1 + p$ coefficients to be estimated (by maximization of the likelihood).

Coefficients

$$\boxed{\begin{array}{l}\text{Pour } j = 1, 2, \ldots, K-1 : \\ \text{logit } \mathbb{P}_\beta(Y \leq j | X = x) = \alpha_j + \beta_1 x_1 + \ldots + \beta_p x_p\end{array}}$$

- Coefficients $(\beta_i)_{i=1}^p$ do not depend on j.
- $\alpha_1 < \alpha_2 < \ldots < \alpha_{K-1}$.
- As $p_\beta^K(x) = 1$, there are $K - 1 + p$ coefficients to be estimated (by maximization of the likelihood).

Interpretation of coefficients

The coefficient β_i measures the conditional dependence between Y and X_i given $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p$.
$\rightsquigarrow$ Nullity of coefficients β

1. Cumulative logit model
   - Generalities
   - Coefficients
2. Estimation, inference
   - Lasso estimation of the coefficients β
   - Penalty parameter and variable selection
3. Simulation studies
4. Application to network inference of zero-inflated data

Lasso optimization

✯ Sparsity assumption to identify the influential covariates (variable selection)

Lasso optimization

✭ Sparsity assumption to identify the influential covariates (variable selection)

✭ Lead to solve the following optimization problem:

$$\underset{\substack{\alpha_1 < \ldots < \alpha_{K-1} \\ \beta \in \mathbb{R}^p}}{\operatorname{argmax}} \left[\log \mathcal{L}_{\alpha,\beta} - \lambda ||\beta||_1 \right]$$

Lasso optimization

★ Sparsity assumption to identify the influential covariates (variable selection)

★ Lead to solve the following optimization problem:

$$\underset{\substack{\alpha_1<\ldots<\alpha_{K-1}\\ \beta\in\mathbb{R}^p}}{\operatorname{argmax}} \left[\log \mathcal{L}_{\alpha,\beta} - \lambda||\beta||_1\right]$$

equivalent to (lagrangian duality):

$$\underset{\substack{\alpha_1<\ldots<\alpha_{K-1}\\ ||\beta||_1\leq\tau}}{\operatorname{argmax}} \log \mathcal{L}_{\alpha,\beta}$$

Frank-Wolfe algorithm

Frank-Wolfe algorithm

$$\min_{x \in C} f(x);\ \ f, C \text{ convex}$$

Iteration k:

- $s^{(k)} \in \underset{s \in C}{\operatorname{argmin}}\, {}^t\nabla f(x^{(k-1)}).s$
- $x^{(k)} = (1-\gamma_k)x^{(k-1)} + \gamma_k s^{(k)},\ \ \gamma_k = \dfrac{2}{k+1}$

☞ linear approximation of the target function

Variable selection

$$\underset{\substack{\alpha_1 < \ldots < \alpha_{K-1} \\ ||\beta||_1 \leq \tau}}{\text{argmax}} \quad \log \mathcal{L}_{\alpha,\beta}$$

☞ How to perform variable selection?

Revisited knockoffs

- Inspired from Barber and Candès (2015)
- Intuitive and suitable to any regression framework
- Provides a sorting of covariates

Revisited knockoffs

- Inspired from Barber and Candès (2015)
- Intuitive and suitable to any regression framework
- Provides a sorting of covariates

Idea

The idea is to use a matrix of knockoffs of covariates whose structure is similar to X but independent from Y:

- If X_i enters the model after its KO $\rightsquigarrow$ X_i does not belong to the model
- Otherwise $\rightsquigarrow$ X_i is more likely to be relevant

Revisited knockoffs

Procedure

1. We construct the knockoffs matrix $\tilde{X}$ by swapping (randomly) the rows of X

Revisited knockoffs

Procedure

1. We construct the knockoffs matrix $\tilde{X}$ by swapping (randomly) the rows of X
2. We calculate statistics $T_i := \inf\ \{\tau > 0,\ \hat{\beta}_i(\tau) \neq 0\}$, $i = 1, \ldots, \mathbf{2p}$ for each variable of the augmented design

Revisited knockoffs

Procedure

1. We construct the knockoffs matrix $\tilde{X}$ by swapping (randomly) the rows of X
2. We calculate statistics $T_i := \inf \{\tau > 0, \ \hat{\beta}_i(\tau) \neq 0\}$, $i = 1, \ldots, \mathbf{2p}$ for each variable of the augmented design
3. For $i \in \{1, ..., p\}$, $W_i := T_i \wedge T_{i+p} \times \begin{cases} +1 & \text{if } T_i < T_{i+p} \\ -1 & \text{if } T_i \geq T_{i+p} \end{cases}$

Revisited knockoffs

Procedure

1. We construct the knockoffs matrix $\tilde{X}$ by swapping (randomly) the rows of X
2. We calculate statistics $T_i := \inf \{\tau > 0,\ \hat{\beta}_i(\tau) \neq 0\}$, $i = 1, \ldots, \mathbf{2p}$ for each variable of the augmented design
3. For $i \in \{1, ..., p\}$, $W_i := T_i \wedge T_{i+p} \times \begin{cases} +1 & \text{if } T_i < T_{i+p} \\ -1 & \text{if } T_i \geq T_{i+p} \end{cases}$
4. Change detection methods applied to the sorted positive statistics W_i to select variables for which statistics W is positive and small

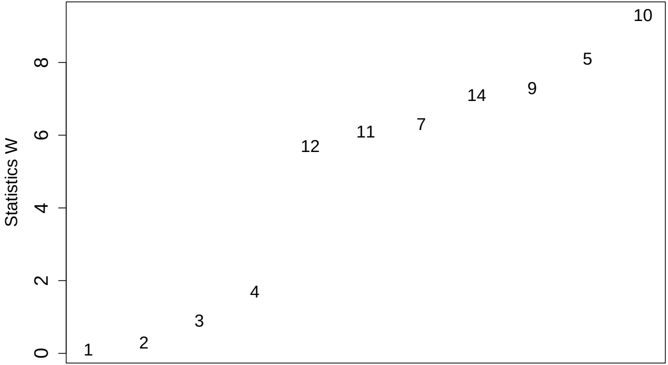

Figure: Example of sorted positive statistics W_i. Only variables X_1, X_2, X_3 and X_4 belong to the model (in this case, $\beta = (8, 6, 4, 2, 0, \ldots, 0)$).

Simulations settings

- Covariates X are simulated as gaussian s.t. X_i conditionaly independent to X_j with probability $1 - p$, $p = 0.6$

Simulations settings

- Covariates X are simulated as gaussian s.t. X_i conditionaly independent to X_j with probability $1 - p$, $p = 0.6$
- $n = 200$ or 100 samples, $p = 50$ covariates

Simulations settings

- Covariates X are simulated as gaussian s.t. X_i conditionaly independent to X_j with probability $1-p$, $p=0.6$
- $n=200$ or 100 samples, $p=50$ covariates
- $K=3$ modalities (for the response variable Y)

Simulations settings

- Covariates X are simulated as gaussian s.t. X_i conditionaly independent to X_j with probability $1 - p$, $p = 0.6$
- $n = 200$ or 100 samples, $p = 50$ covariates
- $K = 3$ modalities (for the response variable Y)
- Regression coefficients $\beta = (8, 6, 4, 2, 0, \dots, 0)$

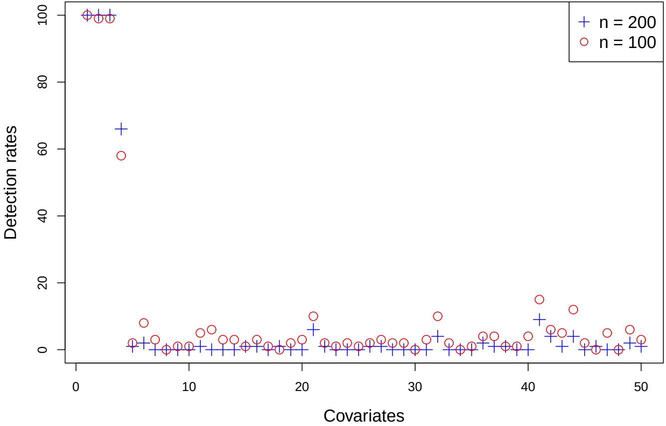

Figure: Detection rates on 100 repetitions after applying revisited knockoffs method.

1. Cumulative logit model
   - Generalities
   - Coefficients

2. Estimation, inference
   - Lasso estimation of the coefficients β
   - Penalty parameter and variable selection

3. Simulation studies

4. Application to network inference of zero-inflated data

Zero-inflated data simulation

Model to simulate data which looks like our kind of data (positive, zero-inflated) and such that we know the theoretical graph structure $\rightarrow$ latent Gaussian graphical model:

Zero-inflated data simulation

Model to simulate data which looks like our kind of data (positive, zero-inflated) and such that we know the theoretical graph structure $\rightarrow$ latent Gaussian graphical model:

- Simulate a Gaussian vector $X \sim \mathcal{N}_p(\mu, \Sigma)$ (we choose the parameters) $\rightsquigarrow$ the graph structure is given by Σ^{-1}.

Zero-inflated data simulation

Model to simulate data which looks like our kind of data (positive, zero-inflated) and such that we know the theoretical graph structure $\rightarrow$ latent Gaussian graphical model:

- Simulate a Gaussian vector $X \sim \mathcal{N}_p(\mu, \Sigma)$ (we choose the parameters) $\rightsquigarrow$ the graph structure is given by Σ^{-1}.
- Simulate a Bernoulli p-random vector such that $Ber_i \sim B(\tilde{p}(X_i))$ for an increasing function $\tilde{p}$.

Zero-inflated data simulation

Model to simulate data which looks like our kind of data (positive, zero-inflated) and such that we know the theoretical graph structure $\rightarrow$ latent Gaussian graphical model:

- Simulate a Gaussian vector $X \sim \mathcal{N}_p(\mu, \Sigma)$ (we choose the parameters) $\rightsquigarrow$ the graph structure is given by Σ^{-1}.
- Simulate a Bernoulli p-random vector such that $Ber_i \sim B(\tilde{p}(X_i))$ for an increasing function $\tilde{p}$.
- Final data are then $Z = \mathrm{Ber} \cdot X$.

Network inference

Goal

Retrieve links of conditional dependence between variables X_i (known in theory thanks to the matrix Σ^{-1}) with the observations Z.

Network inference

Goal

Retrieve links of conditional dependence between variables X_i (known in theory thanks to the matrix Σ^{-1}) with the observations Z.

1. Cumulative logits regression (+ revisited KO) of each variable on the others $\rightsquigarrow$ need to break down the corresponding response variable into classes

Network inference

Goal

Retrieve links of conditional dependence between variables X_i (known in theory thanks to the matrix Σ^{-1}) with the observations Z.

1. Cumulative logits regression (+ revisited KO) of each variable on the others $\rightsquigarrow$ need to break down the corresponding response variable into classes
2. Two networks can be built: the 'and' or the 'or' versions

Simulations settings

- $n = 200$ samples, $p = 200$ variables

Simulations settings

- $n = 200$ samples, $p = 200$ variables
- 'chain' structure

Simulations settings

- $n = 200$ samples, $p = 200$ variables
- 'chain' structure
- zero-inflation: $\approx 12\%$, varying from 0 to 64%

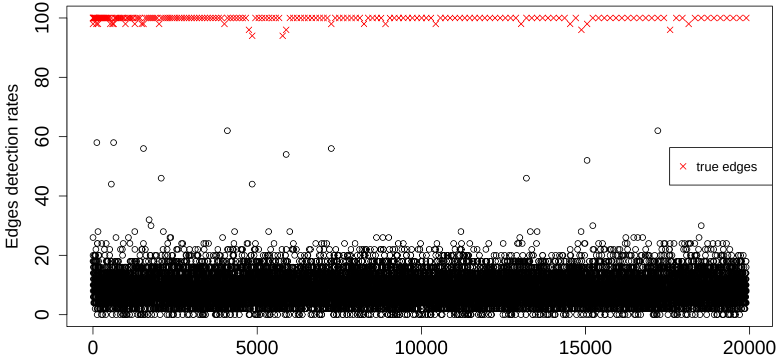

Figure: Edges detection rates (network 'and') on 50 repetitions after applying revisited KO method. Circles and crosses represent respectively false and true edges.

Thank you !